

WHAT IS CLAIMED IS:

1. A method of generating a score for a node identified during a parse of a text segment, the method comprising:

identifying a phrase level for the node;
identifying a word class for at least one word that neighbors a text spanned by the node; and
generating a score based on the phrase level and the word class.

2. The method of claim 1 wherein identifying a word class comprises:

identifying a word class for a word to the left of the text spanned by the node;
and
identifying a word class for a word to the right of the text spanned by the node.

3. The method of claim 2 wherein generating a score comprises generating a score based on the phrase level of the node, the word class of the word to the right of the text spanned by the node and the word class of the word to the left of the text spanned by the node.

4. The method of claim 3 wherein generating a score further comprises determining a mutual information metric.

5. The method of claim 4 wherein determining a mutual information metric comprises determining a mutual information metric based on the phrase level of the node, the word class of the word to the right of the text spanned by the node and the word class of the word to the left of the text spanned by the node.

6. The method of claim 2 wherein identifying a word class further comprises:

identifying all possible word classes for a word to the left of the text spanned by the node; and
identifying all possible word classes for a word to the right of the text spanned by the node.

7. The method of claim 6 wherein generating a score comprises generating a score based in part on all of the identified word classes.

8. The method of claim 1 wherein identifying a word class comprises identifying all possible word classes for at least one word.

9. The method of claim 1 wherein generating a score comprises determining a mutual information metric.

10. A parser for generating a syntax structure from a text segment, the parser comprising:

a seeding unit for inserting words from the text segment into a candidate list as nodes;

a node selector for promoting nodes from the candidate list to a node chart;

a rule engine for combining nodes in the node chart to form a larger node; and

a metric calculator for generating a score for a node formed by the rule engine, the score being based in part on mutual information.

11. The parser of claim 10 wherein the mutual information is determined based on a phrase level of the node formed by the rule engine and at least one word in the text segment.

12. The parser of claim 11 wherein the mutual information is determined based on a word class for a word in the text segment.

13. The parser of claim 12 wherein the mutual information is determined based on all possible word classes for a word in the text segment.

14. The parser of claim 12 wherein the mutual information is determined based on a word class for a word to the left of a set of words spanned by the node formed by the rule engine.

15. The parser of claim 14 wherein the mutual information is determined based additionally on a word class for a word to the right of the set of words spanned by the node formed by the rule engine.

16. The parser of claim 10 further comprising a lexicon look-up for determining parts of speech for words in the text segment.

17. The parser of claim 16 wherein the seeding unit inserts a node for each part of speech of each word in the text segment.

18. The parser of claim 17 wherein the seeding unit further inserts nodes representing the beginning of the text segment and the ending of the text segment.

19. A computer-readable medium having computer-executable instructions for performing steps comprising:

dividing a text segment into words;
forming syntax nodes that each represent a syntax structure for one or more words;
scoring a syntax node to indicate its likelihood of appearing in a full parse structure for the text segment, the score being a mutual information score; and

using the score for the syntax node when forming the full parse structure.

20. The computer-readable medium of claim 19 wherein scoring a syntax node comprises using a mutual information score that is based in part on a phrase level of the syntax node.

21. The computer-readable medium of claim 20 wherein the mutual information score is further based on a word class of a word in the text segment.

22. The computer-readable medium of claim 21 wherein the mutual information score is based on a word class of a word that is next to a word that the syntax node represents.